BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 21 2012, pages 2843–2844
doi:10.1093/bioinformatics/bts521

Sequence analysis Advance Access publication August 24, 2012

HiTC: exploration of high-throughput ‘C’ experiments

Nicolas Servant[1,2,3,*], Bryan R. Lajoie[4], Elphège P. Nora[1,5,6], Luca Giorgetti[1,5,6], Chong-Jian Chen[1,2,3,5,6], Edith Heard[1,5,6], Job Dekker[4] and Emmanuel Barillot[1,2,3]

[1]Institut Curie, F-75248 Paris, France, [2]INSERM, U900, F-75248 Paris, France, [3]Ecole des Mines ParisTech, F-77300 Fontainebleau, France, [4]Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, MA 01605, USA, [5]CNRS UMR3215, F-75248 Paris, France and [6]INSERM U934, F-75248 Paris, France

Associate Editor: Alex Bateman

ABSTRACT

Summary: The R/Bioconductor package *HiTC* facilitates the exploration of high-throughput 3C-based data. It allows users to import and export ‘C’ data, to transform, normalize, annotate and visualize interaction maps. The package operates within the Bioconductor framework and thus offers new opportunities for future development in this field.

Availability and implementation: The R package *HiTC* is available from the Bioconductor website. A detailed vignette provides additional documentation and help for using the package.

Contact: nicolas.servant@curie.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 5, 2012; revised on August 10, 2012; accepted on August 14, 2012

1 INTRODUCTION

The three-dimensional organization of chromosomes and the physical interactions occurring along and between them play an important role in the regulation of gene activity. Over the past 10 years, the development of Chromosome Conformation Capture (3C)-based techniques has changed our view of nuclear organization (de Wit and de Laat, 2012). With the emergence of next-generation sequencing, high-throughput conformation capture techniques, such as Circular 3C (4C; Simonis *et al*., 2006), 3C Carbon-Copy (5C; Dostie *et al*., 2006) or more recently Hi-C (Lieberman-Aiden *et al*., 2009), have been developed to study the physical interactions between many loci in parallel.

While the use of high-throughput ‘C’ techniques is expected to increase in coming years (Dixon *et al*., 2012; Nora *et al*., 2012), bioinformatic methods and software to analyze such data are still lacking. Here, we present the R/Bioconductor package *HiTC* that enables users to visualize and explore high-throughput ‘C’ data. One advantage of the *HiTC* package is that it operates within the open source Bioconductor framework (Gentleman *et al*., 2004) and thus offers new opportunities for future developments in this field. The *HiTC* package is aimed at biologists interested in investigating their data and at biostatisticians involved in the development of new statistical methods which can be applied to C data.

2 AVAILABLE FUNCTIONALITIES

The *HiTC* package provides a variety of functionalities to handle high-throughput C data and is especially suited for visualization and basic transformations of 5C and Hi-C data (Supplementary Fig. S1). Here, we present some of the main functionalities of the package.

2.1 Importing C data

Two distinct datasets are included in the package. The first one is a 5C dataset (Nora *et al*., 2012), corresponding to the X inactivation center obtained in Mouse ES cells (GSE35721) and the second is the Hi-C data from chromosome 14 (GSE18199) published by Lieberman-Aiden *et al*. (2009). Both Hi-C and 5C datasets can be imported using a defined csv format. In addition, the *HiTC* package is fully compatible with data from the my5C web tool (Lajoie *et al*., 2009).

2.2 Quality control

Because of the polymer nature of chromatin, a Hi-C or 5C experiment is expected to be dominated by signal from neighboring restriction fragments in *cis*. Quality control provides simple descriptive statistics and graphical outputs to check the prevalence of *cis*- and *trans*-chromosomal interactions, and to assess whether the expected higher frequency between sites located near each other in the linear genome is verified.

2.3 Visualization

An interaction map is a two-dimensional heatmap representation of the matrix of 5C or Hi-C counts, whose entries correspond to the number of times two restriction fragments in a given genomic region have been ligated in 3C and sequenced as a pair. The *HiTC* package proposes a list of options to define the appropriate data visualization, such as contrast, color or counts trimming. Two different views are provided: a square heatmap view and a triangular view (Fig. 1). The latter is particularly useful for aligning interaction maps and genomic and epigenomic features.

2.4 Interaction map transformation

Depending on the experimental resolution and/or the desired genomic scale to be visualized, each pixel of an interaction map can correspond to a single restriction fragment, several restriction fragments or genomic intervals of any given size (and

[*]To whom correspondence should be addressed.

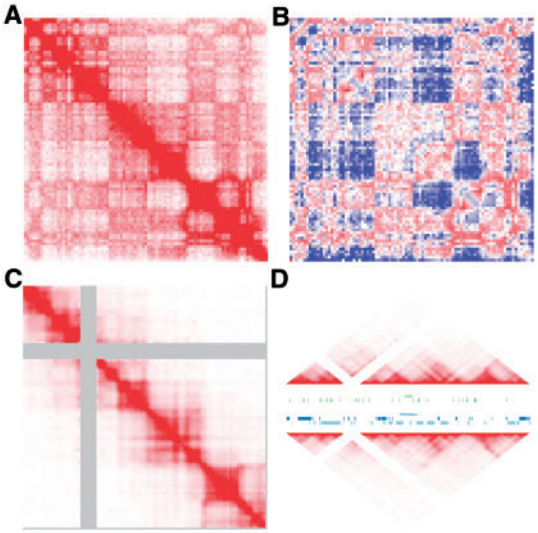

Fig. 1. Visualization of interaction maps. (**A**) Binned interaction map (1 Mb) of Hi-C data (chr14, GSE18199). (**B**) Binned and normalized by expected counts interaction map (1 Mb) of the Hi-C data. (**C**) Heatmap view of the ESC E14 5C interaction map (GSE35721). (**D**) Comparison of the ESC E14 and PGK 5C interaction maps. The genes and CTCF regions from both strands are displayed in blue and green, respectively

therefore various numbers of restriction fragments). For example, 5C allows interaction frequencies to be assessed for each pair of restriction fragments present in the pool of 5C oligonucleotides. The Hi-C protocol, on the contrary, does not necessarily yield counts for every single pair of restriction fragments, especially when analyzing large genomes. Hi-C results are thus typically displayed for genomic bins of an arbitrary size. The *HiTC* package provides a binning function to address the interaction map transformation. For instance, *HiTC* enables the same 5C dataset to be displayed either at the restriction-fragment resolution or after binning in 100 Kb or 1 Mb bins, and these bins can be chosen to partially overlap or not.

2.5 Interaction map normalization

As mentioned earlier, at small genomic distances, pairs of restriction fragments that are close to each other in the linear genome will give higher signal than fragments that are further apart. This leads to most counts mapping to the heatmap diagonal. When considering any given pair of restriction fragments, it can therefore be informative to assess whether the observed counts are above what would be expected given their genomic distance. The *HiTC* package includes a basic normalization function that estimates the interaction counts one would expect if the signal was only dependent on the genomic distance between the interacting loci (Fig. 1B). This calculation is based on Lowess averaging of the observed interaction counts as applied by Bau *et al.* (2011).

3 CONCLUSION

Although we are still far from understanding the exact relationship between chromosome conformation and gene or genome regulations, breakthrough technologies are now available for the systematic and detailed analysis of nuclear organization. The analysis of chromosome conformation capture datasets is quite complex and requires the development of computational tools, including dedicated statistical methods and visualization software, such as the one we propose here. We wish to emphasize that appropriate interpretation of high-throughput C data can require pre-processing of the data, in order to eliminate systematic biases that can be introduced by the experimental protocol or that can arise from the intrinsic properties of the genome, such as a non-homogenous distribution of restriction sites (Yaffe and Tanay, 2011; Zhang and McCord, 2012). The R/BioConductor package *HiTC* proposes a powerful and extensible framework for visualizing and exploring high-throughput C data. It is able to handle both 5C and Hi-C data and offers new functionalities such as standard import, data transformation and integrative visualization methods. While pre-processing and visualization tools started to emerge, other methods such as bias correction, samples comparison or data integration can be further investigated. In this way, the HiTC package provides a flexible basis for further developments by the community.

ACKNOWLEDGEMENTS

The authors thank Joern Toedling, Pierre Gestraud and Marc Carlson for useful discussions and for their help in creating the R package.

Funding: The French program ’investissement d’avenir’ action bioinformatique (ABS4NGS project), the Ministère de la Recherche et de l’Enseignement Supérieur and the ARC (to E.P.N.), EU EpiGeneSys FP7 Network of Excellence no. 257082, the Fondation pour la Recherche Medicale, ANR, ERC Advanced Investigator award no. 250367 and EU FP7 SYBOSS (242129) and MODHEP (259743) (to E.H.). NIH (R01 HG003143) and W.M. Keck Foundation Distinguished Young Scholar Award (J.D. and B.R.L.).

Conflict of Interest: none declared.

REFERENCES

Bau,D. *et al.* (2011) The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. *Nat. Struct. Mol. Biol.*, **18**, 107–114.

de Wit,E. and de Laat,W. (2012) A decade of 3c technologies: insights into nuclear organization. *Genes Dev*., **26**, 11–24.

Dixon,J.R. *et al.* (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376–380.

Dostie,J. *et al.* (2006) Chromosome conformation capture carbon copy (5c): a massively parallel solution for mapping interactions between genomic elements. *Genome Res*., **16**, 1299–1309.

Gentleman,R.C. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*., **5**, R80.

Lajoie,B.R. *et al.* (2009) My5c: web tools for chromosome conformation capture studies. *Nat. Methods*, **6**, 690–691.

Lieberman-Aiden,E. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.

Nora,E.P. *et al.* (2012) Spatial partitioning of the regulatory landscape of the x-inactivation centre. *Nature*, **485**, 381–385.

Simonis,M. *et al.* (2006) Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4c). *Nat. Genet*., **38**, 1348–1354.

Yaffe,E. and Tanay,A. (2011) Probabilistic modeling of hi-c contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat. Genet*., **43**, 1059–1065.

Zhang,Y. and McCord,R.P. (2012) Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. *Cell*, **148**, 908–921.